Neue Ergebnisse aus der KI-Forschung: Menschen können KI-generierte Medien kaum erkennen

Bericht: CISPA Helmholtz Center for Information Security

KI-generierte Bilder, Texte und Audiodateien sind so überzeugend, dass Menschen diese nicht mehr von menschengemachten Inhalten unterscheiden können. Dies ist das Ergebnis einer Online-Befragung mit etwa 3.000 Teilnehmer:innen aus Deutschland, China und den USA. Es ist das erste Mal, dass eine große länderübergreifende Studie, diese Form der Medienkompetenz überprüft hat. CISPA-Faculty Dr. Lea Schönherr und Prof. Dr. Thorsten Holz präsentierten die Ergebnisse diese Woche auf dem 45. IEEE Symposium on Security and Privacy in San Francisco. Die Studie entstand in Kooperation mit der Ruhr-Universität Bochum, der Leibniz Universität Hannover sowie der TU Berlin.

Die rasanten Entwicklungen der letzten Jahre im Bereich der Künstlichen Intelligenz haben zur Folge, dass mit nur wenigen Klicks massenhaft Bilder, Texte und Audiodateien generiert werden können. Prof. Dr. Thorsten Holz erläutert, welche Risiken aus seiner Sicht damit verbunden sind: „Künstlich erzeugter Content kann vielfältig missbraucht werden. Wir haben in diesem Jahr wichtige Wahlen, wie die Wahlen zum EU-Parlament oder die Präsidentschaftswahl in den USA: Da können KI-generierte Medien sehr einfach für politische Meinungsmache genutzt werden. Ich sehe darin eine große Gefahr für unsere Demokratie“. Vor diesem Hintergrund ist eine wichtige Forschungsherausforderung die automatisierte Erkennung von KI-generierten Medien. „Das ist allerdings ein Wettlauf mit der Zeit“, erklärt CISPA-Faculty Dr. Lea Schönherr. „Medien die mit neu entwickelten Methoden zur Generierung mit KI erstellt sind, werden immer schwieriger mit automatischen Methoden erkannt. Deswegen kommt es im Endeffekt darauf an, ob ein Mensch das entsprechend einschätzen kann“. Dies war der Ausgangspunkt um zu untersuchen, ob Menschen überhaupt in der Lage sind, KI-generierte Medien zu identifizieren.

KI-generierte Medien werden mehrheitlich als menschengemacht klassifiziert

Die Ergebnisse der medien- und länderübergreifenden Studie sind erstaunlich: „Wir sind jetzt schon an dem Punkt, an dem es für Menschen schwierig ist – wenn auch noch nicht unmöglich – zu unterscheiden, ob etwas echt oder KI-generiert ist. Und das gilt eben für alle Arten von Medien: Text, Audio und Bild“ erklärt Holz. Die Studienteilnehmer:innen klassifizierten KI-generierte Medien über alle Medienarten und Länder hinweg mehrheitlich als menschengemacht. „Überrascht hat uns, dass es sehr wenige Faktoren gibt, anhand derer man erklären kann, ob Menschen besser im Erkennen von KI-generierten Medien sind oder nicht. Selbst

über verschiedene Altersgruppen hinweg und bei Faktoren wie Bildungshintergrund, politischer Einstellung oder Medienkompetenz, sind die Unterschiede nicht sehr signifikant“, so Holz weiter.

Medien-Erkennung mit Abfrage soziobiografischer Daten kombiniert

Die quantitative Studie wurde als Online-Befragung zwischen Juni 2022 und September 2022 in China, Deutschland und den USA durchgeführt. Per Zufallsprinzip wurden die Befragten einer der drei Mediengruppen „Text“, „Bild“ oder „Audio“ zugeordnet und sahen 50% reale und 50% KI-generierte Medien. Darüber hinaus wurden sozio-biografische Daten, das Wissen zu KI-generierten Medien sowie Faktoren wie Medienkompetenz, holistisches Denken, generelles Vertrauen, kognitive Reflexion und politische Orientierung erhoben. Nach der Datenbereinigung blieben 2.609 Datensätze übrig (822 USA, 875 Deutschland, 922 China), die in die Auswertung einflossen. Die in der Studie verwendeten KI-generierten Audio- und Text-Dateien wurden von den Forscher:innen selbst generiert, die KI-generierten Bilder übernahmen sie aus einer existierenden Studie. Die Bilder waren fotorealistische Porträts, als Texte wurden Nachrichten gewählt und die Audio-Dateien waren Ausschnitte aus Literatur.

Ausgangspunkte für weitere Forschung

Das Ergebnis der Studie liefert wichtige Take-Aways für die Cybersicherheitsforschung: „Es besteht das Risiko, dass vor allem KI-generierte Texte und Audio-Dateien für Social Engineering-Angriffe genutzt werden. Denkbar ist, dass die nächste Generation von Phishing-Emails auf mich personalisiert ist und der Text perfekt zu mir passt“, erläutert Schönherr. Abwehrmechanismen für genau solche Angriffsszenarien zu entwickeln, darin sieht sie eine wichtige Aufgabe für die Zukunft. Aber aus der Studie ergeben sich auch weitere Forschungsdesiderata: „Zum einen müssen wir besser verstehen, wie Menschen überhaupt noch KI-generierte Medien unterscheiden können. Wir planen eine Laborstudie, wo Teilnehmer:innen uns erklären sollen, woran sie erkennen, ob etwas KI-generiert ist oder nicht. Zum anderen müssen wir überlegen, wie wir das technisch unterstützen können, etwa durch Verfahren zum automatisierten Fakt-Checking,“ so Schönherr abschließend.

Originalpublikation:
Joel Frank, Franziska Herbert, Jonas Ricker, Lea Schönherr, Thorsten Eisenhofer, Asja Fischer, Markus Dürmuth and Thorsten Holz. “A Representative Study on Human Detection of Artificially Generated Media Across Countries.” ArXiv abs/2312.05976 (2023) *https://doi.org/10.48550/arXiv.2312.05976*

21.5.2024
Felix Koltermann Unternehmenskommunikation
CISPA Helmholtz Center for Information Security
www.cispa.de